Bolt Beranek and Newman Inc.

(1)

REPORT No. 4061

SPEECH COMPRESSION AND SYNTHESIS

QUARTERLY PROGRESS REPORT No. 3
6 OCTOBER 1978 - 5 JANUARY 1979

PREPARED FOR:
ADVANCED RESEARCH PROJECTS AGENCY

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER BBN Report No. 4061	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER 6 October 78 - 5 Jan 79
4. TITLE (and Subtitle) SPEECH COMPRESSION AND SYNTHESIS		5. TYPE OF REPORT & PERIOD COVERED Quarterly Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lynn Cosell, John Makhoul, A.W.F. Huggins, Richard Schwartz, John Klovstad, Jared Wolf		8. CONTRACT OR GRANT NUMBER(s) F19628-78-C-0136
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Massachusetts 02138		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy for Electronic Technology (RADC/ETC) Hanscom Air Force Base, MA 01731 Contract Monitor: Mr. Caldwell P. Smith		12. REPORT DATE 13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION IS UNLIMITED (A)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 3515.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech synthesis, phonetic synthesis, diphone, LPC synthesis, vocoder, speech compression, linear prediction, voice-excited coder, high-frequency regeneration.

(Minor Revision to Cat 2)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This document reports progress in the development of a phonetic speech synthesis algorithm, implementation and development of a real-time LPC vocoder. Testing of spectral modeling using adaptive lattice methods, and results of a subjective evaluation of the mixed source excitation in LPC synthesis. A new diphone utterance data base has been designed and is being recorded for the phonetic synthesis program.

Keywords include: Voice-excited coder and high-frequency regeneration.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 55 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

SPEECH COMPRESSION AND SYNTHESIS

Quarterly Technical Progress Report No. 3

6 October 1978 - 5 January 1979

ARPA Order No. 3515

Contract No. F19628-78-C-0136

Name of Contractor:
Bolt Beranek and Newman Inc.

Principal Investigators:
Dr. John Makhoul
(617)491-1850, x4332

Dr. R. Viswanathan
(617)491-1850, x4336

Effective Date of Contract:
6 April 1978

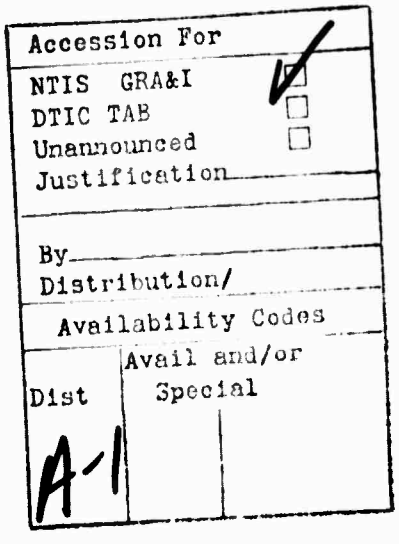
Contract Expiration Date:
5 April 1979

Sponsored by

Defense Advanced Research Projects Agency (DoD)

Monitored by RADC/ETC

TABLE OF CONTENTS



1. SUMMARY

Below is a summary of our progress in the third quarter of this contract. The details are given in Sections 2-5.

1.1 Phonetic Synthesis

During the end of the second quarter and the beginning of the third quarter, we tried to evaluate the performance of different parts of the phonetic synthesis program. The program could substitute an appropriate module from an LPC vocoder for each synthesis module, thereby allowing us to isolate the effects of each synthesis module. A few general problems were underscored by this effort -- in particular we became aware that we need to pay careful attention to the handling of the gain parameter.

The results of these experiments convinced us that it was necessary to redesign and rerecord the diphone data base in order to achieve high quality. Much of the past quarter has been spent in this redesign and rerecording.

During the initial phase of the project we came to realize that manually segmenting the data base would be a very lengthy process. This was due, in part, to the nature of our existing programs for this task. Therefore some effort was spent in improving these programs and adding others that would speed up the transcription.

1.2 Spectral Estimation

We have begun research into using the adaptive lattice method of spectral estimation to improve the spectral estimate achieved by linear prediction. Several algorithms for using this method are being investigated.

1.3 Real Time Vocoder

An error in the real-time buffering in the implementation of the ARPA network vocoder was detected and fixed. This error had been causing "pops" in the vocoded speech. We are also looking at problems in the BBN configuration arising with conferencing.

We have tested the new Variable Frame Rate (VFR) parameter transmission algorithm. The result is that at data rates around 2200 bits per second (and below) the quality of speech from the new algorithm sounds significantly better than that produced by the older algorithm at the same data rates.

1.4 Subjective Quality Evaluation of Mixed Source Excitation Model

A formal subjective quality evaluation of our new mixed source model was performed. The test confirmed that listeners perceived the speech synthesized using this model to be less buzzy and generally of higher quality than speech produced using a binary pulse/noise excitation model.

2. SPEECH SYNTHESIS

2.1 Algorithm Development

The phonetic speech synthesis program we have developed accepts as input a sequence of triplets. Each triplet consists of a phoneme, a phoneme duration, and a single pitch value. These triplets are determined from the "target sentence" (an actual utterance of the sentence being synthesized). The output of the phonetic synthesis program is a set of parameters for an LPC synthesizer. The LPC synthesizer requires a complete set of "synthesis parameters". They are, for each 10 ms, 14 Log Area Ratio (LAR) parameters specifying a spectrum, a value of gain, a voicing flag, and (if voiced) a value of pitch and cutoff frequency for the mixed source model.

In converting the input sequence of triplets to the synthesis parameters, the LAR parameters and gain are determined from diphone templates that have been extracted from the carefully constructed data base of diphone utterances. The voicing flag and cutoff frequency are determined from the phoneme identity by rule, and the pitch values are derived from the single input values by linear interpolation. The templates are time warped so that the durations of the resulting phonemes match the durations specified in the input string.

As an aid in evaluating the performance of the individual synthesis algorithms, the synthesis program contains provision for specifying the source of any of the synthesis parameters as natural (directly from the target sentence) or synthetic (from the synthesis algorithms). The source for each (of the 5 types) of synthesis parameters can be varied independently by setting appropriate flags, making it possible to investigate the effect of determining each parameter by the synthesis procedure. Of course, if all the synthesis parameters are taken directly from the target sentence, the program becomes an LPC vocoder. Some of the more interesting results obtained from experiments of this type are given in the remainder of this section.

There was no noticeable difference in speech quality when the synthesis program used the target sentence pitch directly instead of the linearized pitch (calculated by interpolating between the single pitch values given as input). This significant result tells us that, for English, a model of pitch that is linear within each phoneme is adequate.

Determining the voicing flag from the identity of the phoneme being synthesized eliminated the voicing errors made by the analysis program. addition, the voicing sounded correct even when this algorithm gave results somewhat different from a "correct" voiced/unvoiced decision. For example, even though

voiced consonants are often devoiced in natural speech, the speech will always sound natural if voiced consonants are always all voiced (with an appropriate cutoff frequency). Using the cutoff frequency derived by rule from the phonemes resulted in more natural sounding speech than that obtained with the cutoff derived from the target sentence.

Our initial attempts at using the LAR (spectral) parameters from the templates in conjunction with all other synthesis parameters taken from the target sentence were somewhat disappointing, since the resulting speech contained some pops and other disturbances. We thought we were not adequately reproducing appropriate LAR parameters by time-warping, concatenating, and smoothing the sets of diphone template data. However, most of these disturbances were eliminated when the gain was also extracted from the templates instead of the target sentence. This last result led us to believe that a major problem was inexact timing between the LAR and gain parameters. To test this hypothesis we tried combining the LAR parameters from the target sentence with the gain from the templates. As expected, this also resulted in many discontinuities and lesser quality than when both come from the templates.

After listening to several such combinations, we concluded that a major remaining factor limiting the naturalness of the

speech (synthesized using both gain and LAR parameters from the templates) was the incompatibility of the gain values in adjacent diphones. This incompatibility was due to the diphones being abutted, having been extracted from differing types of context. This incompatibility was one of the major reasons for redesigning and rerecording the data base of diphone utterances.

2.2 New Data Base

During this quarter, we have redesigned and begun to re-record our data base of diphone utterances. There were several factors that made this advisable. As mentioned above, a significant problem with the first data base was that the gain parameter was not consistent between diphones that were to be abutted to each other. This was partially due to the fact that some of the diphones were taken from the beginning of an utterance, while others were taken from the middle or end of an utterance. In addition, different diphones containing the same vowel were recorded several minutes or even hours apart, and incidental changes in speaking level and mouth-to-microphone distance produced noticeable differences in amplitude. However, there was no consistent method to determine whether one diphone was louder than another simply because it was inherently louder, or because the speaker just happened to be speaking louder.

Since there is evidence that gain was a major problem with the synthesized speech, the recording procedure was designed so that incidental differences in loudness could be minimized and compensated. The utterances were reorganized into groups according to the vowel phoneme of each diphone. Thus, the different diphones involving one particular vowel were all spoken within a short period (roughly 2 minutes). This close proximity helps to ensure that the speaking level was roughly constant during different instances of the same vowel.

The diphone utterances consisted of short nonsense syllables which were repeated three times as in connected speech (e.g. [pa pa pap]). The diphones are to be extracted from the middle syllable, which tends to have a more prototypical articulation (and loudness).

In addition to short-term variations in speaking level, there is also a long-term variation in the level over several hours and between the several days that elapsed during the recordings. In order to estimate this effect, each group of utterances was initiated and terminated with a "normalization utterance". The normalization utterance we chose to use was [dædædæd]. This utterance was chosen, because the vowel [æ] in combination with a voiced plosive results in a higher amplitude than most syllables. After the data has been analyzed, the level of each diphone can be

set relative to the normalization utterances, thus cancelling out long-term variations in speaking level. This framework also allows for the level adjustment that will probably be necessary for several groups of diphones.

The data base contains utterances for all the diphones that are felt to result in different acoustic patterns. The types of diphones included are shown below.

C stands for consonant; V stands for vowel

DIPHONE	EXAMPLE
CV	[pa] as in "pot"
VC	[ap] as in "top"
initial cluster	[spr] as in "spring"
final cluster	[nd] as in "and"
CC	[sf] as in "this formant"
VV	[iæ] as in "reality"

In addition to the vowels, we have included several other vowel allophones such as retroflexed vowels (vowels followed by [r]), lateralized vowels (vowels followed by [l]), [ə] (as in "about"), ɪ (as in "multiply"), syllabic nasals, and syllabic [l]. The consonants include silence, flapped [t], unreleased plosives, affricates, and glottal stops. Due to the inclusion of all

permutations of these phonemes, the new recording includes a total of 1894 utterances representing 3145 diphones. However, of the 3145 diphones, several cannot occur in English, and many are likely not to be necessary as separate diphones. We will use as many of these diphones as are necessary to achieve natural sounding speech. We expect that this will require approximately 2500 diphones.

Since these recordings were to form the basis for the synthesis to be done in the remainder of this project, the recordings were monitored carefully. It was felt that, for this application, very low noise recordings were desirable. The recordings were made in a quiet room. The microphone used was an "electret" condenser microphone positioned 2 inches from the right corner of the mouth at an angle of 45 degrees to the side. The close-talking microphone was chosen over a distant (approx. 11 inches) microphone because it allowed us to attenuate low level building-borne noise. Also, the quality of the microphone was judged to be quite high. The recordings are being made on a Braun TG-1000 tape deck.

Each utterance (including the two normalization utterances for each group of diphones) is digitized into a single speech file using 12 bits per sample. The dynamic range of most of the utterances only requires 11 bits.

In order to extract the diphones from these relatively long (approximately 1 second) diphone utterances, we examine the utterance and indicate the identity and end points of the relevant phonemes. This transcription must be accurate since the spectra and energy parameters derived from the speech will be combined with voicing information determined from the phoneme identity. Therefore misalignment will result in apparent "voicing" errors.

2.3 Display Programs

As previously mentioned, a significant part of this project consists of accurately and consistently transcribing a complete set (approximately 2500) of diphones. This task can be more efficiently accomplished on the PDP-11 system than on the PDP-10, since the PDP-11 is equipped with a digital playout system and can utilize the AP120B array processor for very high speed signal processing. The PDP-11 is also linked to the IMLAC PDS-1 graphics display terminal via a very high speed parallel connection.

Our real-time waveform editing program was modified so that, in addition to displaying, editing and playing time waveforms, it could compute and display the instantaneous spectrum (log-magnitude power spectrum and LPC spectral envelope) corresponding to a short window of speech pointed to by a cursor on the waveform display. As the cursor is moved relative to the waveform, the program

recomputes and displays the spectra (corresponding to the instantaneous cursor location) eight times per second.

The program also allows interactive manual transcription of time waveforms. The display of the speech waveform and the short term spectra are sufficient indicators for the majority of phoneme boundaries. Those phonemes that require more abstract representations, such as formant tracks, will be transcribed on the PDP-10 using existing (but slower) software. We estimate that this addition to the PDP-11 facility has halved the time that will be necessary for the manual transcription. The program will also be useful in other future applications.

3. ADAPTIVE LATTICE METHOD OF PARAMETER ESTIMATION

In an attempt to improve the reliability of spectral estimation we began to investigate adaptive lattice analysis. It is our hope that this analysis method will provide, in effect, pitch synchronous estimation.

The adaptive lattice analysis method used is described in reference [1]. The reflection coefficients, K_m, are estimated at each sampling interval, n, as shown:

$$K_m(n+1) = -\frac{2\sum_{k=-\infty}^{n} w(n-k)\, f_{m-1}(k)\, g_{m-1}(k-1)}{\sum_{k=-\infty}^{n} w(n-k)\, [f_{m-1}^2(k) + g_{m-1}^2(k-1)]}$$

$$= -\frac{C_m(n)}{D_m(n)}$$

where w(n) is the window function on the error signal, and $f_m(n)$ and $g_m(n)$ are given by:

$$f_0(n) = g_0(n) = s(n)$$

$$f_m(n) = f_{m-1}(n) + K_m g_{m-1}(n-1)$$

$$g_m(n) = K_m f_{m-1}(n) + g_{m-1}(n-1)$$

where s(n) is the sampled speech signal.

Since the adaptive lattice method generates a new set of reflection coefficients every sample time, our research has been

directed toward determining the best set to use for transmission. Our first approach was to low-pass filter $K_m(n)$, and then sample the filtered versions once each transmission interval. This approach had the unfortunate side effect of muddying the resultant synthetic speech, particularly when the speech is changing rapidly.

We have also begun to investigate the extent of variation in the set of reflection coefficients during a transmission interval. We have compared the reflection coefficients generated by the adaptive lattice with those produced by the usual autocorrelation linear prediction method when the analysis window is moved along in time in one sample increments. The autocorrelation method produces reflection coefficients that vary only slightly across the analysis interval. However, when the autocorrelation method is used with a rectangular window instead of the normal Hamming window, the variation is greatly increased and this variation is correlated with the pitch pulse. Using the lattice method, we also see variations in the reflection coefficients that are highly correlated with the pitch pulses. We have found that the shape of error weighting window w(n) has a large effect on the magnitude of these variations. For a recursive window that is the impulse response of a single real pole $w(n)=\beta^n$, $0<\beta<1$, we found that increasing β reduces the magnitude of the variation. However, increasing β also tends to smear the resultant synthetic speech.

Using a double-pole window, $w(n)=(n+1)\beta^n$, allows a greater reduction in the variation coupled with less smearing.

During the next quarter we plan to continue our investigation of adaptive lattice analysis, in order to find criteria that enable us to select the most reliable set of reflection coefficients. We will continue to investigate the effect of different error weighting windows.

4. REAL-TIME VOCODER

We have managed to find and fix an annoying problem in the implementation of the real-time vocoder. This problem, manifested by pops and clicks in the output speech, was discovered to be a real-time buffering problem. The solution avoids the pops and clicks, and at the same time provides a little more elasticity in the time constraints on processing a buffer of data. We have been working closely with ISI in finding and fixing the problem.

We (and ISI) have also made progress in identifying and relieving our difficulties with conferencing. Currently the BBN configuration cannot participate in a conference satisfactorily. We will continue experimenting with conferencing in the next quarter.

We have performed a preliminary evaluation of the speech quality produced by the real-time vocoder using the new Variable Frame Rate (VFR) (optimal-linear-fit) algorithm. We have found, using RTFUD, that the vocoder using the older form of VFR transmission transmits continuous speech at a rate of about 2800 bits/second, using the nominal value for the VFR threshold. At a transfer rate this high, we hear little, if any, quality improvement with the new VFR algorithm. However, for transmission rates of about 2200 bits/second, the new method produces speech

that is clearly superior to that produced by the old method. The quality produced by the old method degrades rapidly as the transmission rate is reduced beyond a certain point (approximately 2400 bits/second for continuous speech). The new method degrades more slowly, and continues to degrade slowly even when the transmission rate is reduced beyond this point. Since the transmission rate and the quality of the speech depend on the specific utterance, the bit rates discussed are approximate and should be understood to be only indicative of a general trend.

5. SUBJECTIVE QUALITY EVALUATION OF MIXED SOURCE EXCITATION MODEL

We performed a formal subjective quality evaluation of our new mixed source model, described in earlier QPRs. The test had four purposes:

1) to confirm that the mixed source model yields better speech quality than the conventional binary excitation model;

2) to demonstrate that remaining weaknesses in the mixed source model can be ascribed to the algorithm for selecting the filter cutoff frequency defining the boundary between pulse and noise excitation, rather than to the model itself;

3) to show that raising the cutoff frequency above that specified by the algorithm tends to both increase the buzziness and reduce the breathiness of the speech, while lowering the cutoff has the reverse effects;

4) to identify test sentences on which the mixed source model yielded less improvement in quality than expected, for the purpose of later using these sentences for improving the algorithm.

5.1 Mixed Source Model

Most current LPC vocoders specify each frame of speech as either voiced or unvoiced. Voiced frames are excited with a pulse source during resynthesis, and unvoiced frames are excited with a noise source. Our mixed source model "softens" the binary decision by exciting each nominally voiced frame with a mixture of low-pass filtered pulses and high-pass filtered noise, as indicated in Figure 1. The low- and high-pass filters have the same cutoff frequency and roll-off, with the result that their combined output still has a flat spectrum envelope. Thus the conventional binary distinction between voiceless and voiced categories is replaced by a multi-level distinction between voiceless and several degrees of voicedness, specified by a nominally continuous variable, F(c), corresponding to the cross-over frequency between pulse and noise excitation. Theoretically, F(c) can vary over the bandwidth of the speech, but in our implementation we limited it to 9 levels between 500 Hz and 4.5 kHz, in 500 Hz steps. The cutoff frequency is determined, during analysis, by finding the highest frequency region, at least 600 Hz wide (or 2.2 times the pitch frequency if this is larger), in which adjacent peaks in the signal spectrum are separated by a frequency roughly equal to the fundamental. First order Butterworth filters were used for low-passing the pulses and high-passing the noise. More details of the model and its implementation can be found in earlier reports and in [2].

5.2 Experimental Design

The constraints under which we designed the test were:

a) that the conventional (binary) source model and the mixed source model be compared

b) under conditions where other sources of degradation were minimized

c) using a wide range of speech materials

d) spoken by talkers with a wide range of characteristics.

The latter two constraints are well met by the set of six phoneme-specific sentences, read by six selected speakers, that we developed earlier in the contract. The sentences were as follows:

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea-party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

We compared the conventional binary voicing model (BIN) with five different versions of the mixed source model. One of these (MIX-0) was the version specified above and in our earlier work. The other four versions used exactly the same analysis procedures as MIX-0, but before resynthesis the cutoff frequency between pulse

and noise excitation was moved down by two steps (MIX-2), or by one step (MIX-1), or raised by one step (MIX+1) or by two steps (MIX+2), each step corresponding to 500 Hz. One exception to this rule was that, although nominally voiced frames were allowed to become unvoiced, by having their cutoff frequency moved to 0 Hz, the reverse was never allowed: no unvoiced frame became voiced. The reason for evaluating five different versions of the mixed source model was to try to determine whether any remaining distortion due to the mixed source model could be ascribed to the algorithm for extracting the cutoff frequency, or whether, perhaps, the mixed model itself could be the cause.

To minimize extraneous sources of degradation, we selected a single high quality LPC vocoder, and excited it by each of the six source models. The high quality system had 14 poles, and a (fixed) frame rate of 100 frames/sec. The predictor coefficients were not quantized. Each of the thirty-six test sentences (6 speakers x 6 sentences) was processed by the system excited by each of the six source models in turn (BIN, MIX-2, MIX-1, MIX-0, MIX+1, and MIX+2), yielding a total of 216 stimulus sentences. These were recorded on tape in blocks of 6, in two carefully counterbalanced orders in which each speaker, sentence, and system occurred once in every block, and each speaker, sentence, and system followed each other speaker, sentence, and system with the same frequency, to

counterbalance sequential effects on judgments. At the start of each tape, two blocks of six familiarization stimuli were played, in which a single sentence was played through a contrasting pair of systems. Each speaker and sentence was represented once in the familiarization blocks. In addition, the first six blocks of test stimuli on each tape were repeated identically at the end, to permit any drift or lack of repeatability in the subjects' performance to be detected.

5.3 Subjects and Task

Eight subjects served, four of whom were highly experienced in listening to vocoded speech, and four were naive. All were native speakers of English, and reported no hearing difficulties. Each subject made three passes through the 216 stimuli. On the first pass, all subjects judged overall quality of the stimuli, on an 8-point scale (1-8) with "overflow bins" of 0 and 9, which were to be used only if a more extreme stimulus followed one that had been labelled 1 or 8. On the second and third passes, subjects rated each stimulus for buzziness on one pass, and for breathiness on the other. Since these are unipolar rather than bipolar scales, subjects were told to assign 1's to sentences that showed no buzziness (or breathiness), and that they could use as many of the eight points of the scale as they liked. Most subjects used the range 1-5 for all but a few stimuli.

In preliminary listening, we noticed an appreciable incidence of clicks in some of the test sentences. To try to track down their source, we gave subjects a secondary task, telling them to carry it out only if doing so would not interfere with the primary rating task. The secondary task was to write a "C" by the rating if the stimulus sentence was noticeably clicky.

Since there were only two experimental tapes, each subject heard one tape twice, although subjects were not aware of the repetition. The assignment of tapes to conditions, and the order of tasks (except for overall quality judgments, which were always first) were counterbalanced. Subjects ran themselves individually, listening through high quality headphones, and were asked to make all their judgments on one tape without stopping the tape. Interstimulus interval was five seconds for the first few blocks, and decreased to 3.5 seconds thereafter. Interblock intervals were 5 seconds, and a longer pause occurred after every six blocks. Each session took about 30 minutes.

5.4 Results

5.4.1 Reliability

The first six blocks of stimuli heard by each subject on each pass were repeated identically at the end of the pass, without the subject's knowledge. Pearson-R correlation coefficients were

calculated for each pass by each subject, using these 36 repeated judgments, and then the first set of judgments were discarded. The correlations are shown below in Table 1: all correlations were greater than 0.85, and above 0.9 for the quality judgment pass which was the most important. (A coefficient of 0.554 is statistically significant at P<0.001 with 30 degrees of freedom.) Thus, each subject demonstrated a high level of reliability in his/her judgments.

SUBJECT	QUALITY	BUZZINESS	BREATHINESS
1	0.9688	0.8848	0.9434
2	0.9320	0.9232	0.9151
3	0.9519	0.9364	0.9204
4	0.9373	0.9593	0.9370
5	0.9813	0.9461	0.9310
6	0.9098	0.9038	0.8932
7	0.9665	0.9288	0.9009
8	0.9248	0.8501	0.8661

Table 1: Correlation coefficients between replicated first six blocks, for each pass by each subject. A correlation coefficient of 0.554 is significant at P<0.001.

5.4.2 Quality Judgments

Analysis of variance of the quality ratings showed that the main effects of speaker, sentence, and source model were all statistically highly significant, as also were the two and three-way interactions between them. There was no difference between experienced and naive subjects. The average ratings of overall quality as a function of the excitation appear at top of

Figure 2. The mean quality ratings are given for each speaker, sentence, system, and subject, in Table 2. The plots show that MIX-0, MIX+1, and MIX+2 all yielded better quality than BIN, the conventional binary model. This was true even though all the systems were of high quality, which might be expected to mask differences between them due to ceiling effects. Moreover, the differences between the systems were statistically highly significant, as can be seen from the results of pairwise t-tests shown in Table 3. The t-values correspond closely to the number of standard deviations for the normal distribution, when the number of degrees of freedom exceeds 200, as it does in this case. A t-value of 1.96 is statistically significant at $P<0.05$, and a value of 3.29 is significant at $P<0.001$. Hence it can be seen that the differences between MIX-0, MIX+1, and MIX+2 are not statistically significant (bottom three entries in top half-matrix), but that each of these three systems outperformed the binary system with a confidence level of better than 1 in a hundred million.

One mixed source system, MIX-2, yielded significantly worse quality than BIN, and we next turn to the results of the "clickiness" judgments to discover why.

SPEAKER:		AR(f)	JB(m)	DK(m)	DD(m)	RS(f)	PF(f)
SPEAKERS	AV:	3.93	4.81	4.58	4.10	5.26	4.58
(N= 288)	SD:	1.62	1.70	1.83	1.54	1.71	1.65

SENTENCE:		Away	Nanny	His	Party	Little	Swim
SENTENCE	AV:	4.59	4.65	3.84	4.53	5.06	4.59
(N= 288)	SD:	1.87	1.68	1.47	1.66	1.84	1.63

SYSTEM:		BIN	MIX-2	MIX-1	MIX-0	MIX+1	MIX+2
SYSTEMS	AV:	4.20	4.08	4.55	4.80	4.82	4.82
(N= 288)	SD:	1.75	1.82	1.64	1.69	1.68	1.67

SUBJECTS:		1	2	3	4	5	6	7	8
SUBJECTS	AV:	5.46	3.81	4.67	4.12	4.49	4.18	5.09	4.53
(N= 216)	SD:	1.32	1.55	1.80	2.03	1.08	1.67	2.02	1.39

SYSTEM:	BIN	MIX-2	MIX-1	MIX-0	MIX+1	MIX+2	B	-2	-1	-0	+1	+2
SPKR 1 = AR	3.10	4.27	4.35	4.02	3.85	3.96	.0	.0	.0	.0	.0	.0
SPKR 2 = JB	4.23	4.25	4.88	5.04	5.06	5.38	.0	.2	.0	.1	.1	.0
SPKR 3 = DK	4.31	3.48	5.06	4.75	4.94	4.96	.0	.4	.0	.0	.0	.0
SPKR 4 = DD	3.94	3.75	4.06	4.17	4.50	4.17	.1	.4	.2	.1	.0	.1
SPKR 5 = RS	5.19	4.63	4.63	6.04	5.56	5.54	.1	.1	.0	.0	.0	.0
SPKR 6 = PF	4.46	4.10	4.31	4.75	4.98	4.90	.1	.2	.1	.0	.1	.1

SYSTEM:	BIN	MIX-2	MIX-1	MIX-0	MIX+1	MIX+2	B	-2	-1	-0	+1	+2
SENT1=Away	4.69	3.88	4.19	4.92	4.85	5.04	.0	.1	.1	.0	.0	.0
SENT2=Nanny	4.10	4.48	4.73	4.94	4.81	4.83	.1	.2	.1	.1	.1	.1
SENT3=His	3.00	3.71	4.08	4.10	4.08	4.04	.1	.3	.1	.1	.1	.1
SENT4=Party	4.50	3.42	4.42	4.08	5.06	4.92	.0	.3	.0	.0	.0	.0
SENT5=Little	4.71	4.73	5.04	5.29	5.42	5.19	.0	.2	.0	.0	.0	.0
SENT6=Swim	4.23	4.27	4.83	4.65	4.67	4.88	.0	.1	.0	.0	.0	.0

Table 2: Mean quality judgments by speaker, sentence, system, and subject (top), and by speaker and system, and by sentence and system (bottom). The figures at the right give the proportion of occasions on which a system was judged clicky, as a function first of the speaker, and second of the sentence.

QUALITY

	MIX-2	MIX-1	MIX-Ø	MIX+1	MIX+2
BIN	Ø.91	-2.81	-5.73	-6.25	-6.19
MIX-2		-4.71	-5.80	-6.01	-5.97
MIX-1			-2.51	-2.70	-2.54
MIX-Ø				-Ø.23	-Ø.23
MIX+1					Ø.00

BUZZINESS

	MIX-2	MIX-1	MIX-Ø	MIX+1	MIX+2
BIN	13.24	11.24	9.01	7.64	7.57
MIX-2		-3.50	-6.65	-9.32	-8.68
MIX-1			-3.93	-6.42	-5.81
MIX-Ø				-2.89	-2.55
MIX+1					Ø.21

BREATHINESS

	MIX-2	MIX-1	MIX-Ø	MIX+1	MIX+2
BIN	-11.02	-6.92	-3.57	-2.32	-1.50
MIX-2		6.08	8.84	10.23	10.69
MIX-1			4.60	5.95	6.40
MIX-Ø				1.60	2.44
MIX+1					Ø.89

Table 3: Results of t-tests performed on ratings of quality, buzziness, and breathiness, between all pairs of systems. With 212 degrees of freedom, a t-value of 1.96 is significant at P<Ø.Ø5, one of 2.58 is significant at P<Ø.Ø1, and one of 3.29 is significant at P<Ø.ØØ1. Negative values of t indicate that the column system displayed more of the measured attribute than the row system.

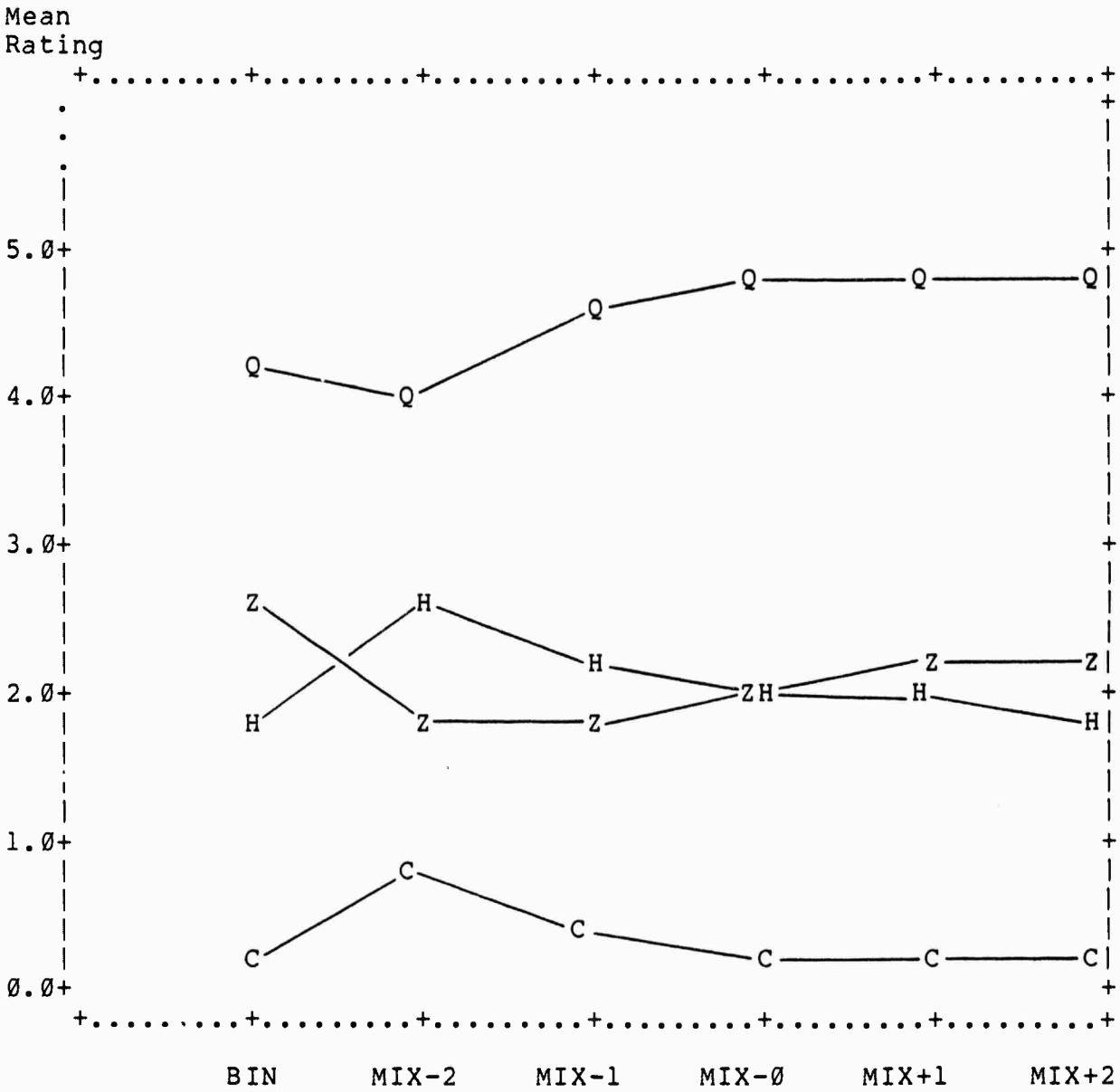


Figure 2: Plots of mean rated quality (Q), buzziness (Z), breathiness (H), and relative incidence of "click" reports (C: 100% = 4.0), for the binary (BIN) and five versions of the mixed (MIX) excitation models. The absolute values of the four different lines are not related -- therefore, for example, there is no significance in the fact that the buzziness and breathiness lines cross at MIX-0.

5.4.3 Clicks

Not all subjects found they were able to make judgments on clickiness on all three passes. In fact, clicks were marked on only 18 out of the total of 24 passes. Of the total of 278 clicky judgments assigned during these 18 passes, more than 47% fell in sentences processed by MIX-2, with over half of these occurring in just six test sentences. As for the other systems, MIX-1 accounted for a further 17%; BIN accounted for 12%; and MIX-0, MIX+1, and MIX+2 accounted for 8% each. The six sentences most severely affected were all either from male speakers (with low fundamental frequencies), or contained a heavy loading of fricatives or stops.

The BIN system obtained better quality ratings than any MIX system on only three out of the 36 test sentences, and two of these were sentences in which MIX-2 and MIX-1 had yielded a high rate of clicky judgments. The margin of superiority in the remaining sentence was very small. In a direct comparison between BIN and MIX-0, MIX-0 was judged superior on 25 test sentences, equal on a further 4, and inferior to BIN on 7. The conclusion seems warranted that the mixed source model is inherently superior to the binary model, and that any remaining weaknesses it displays are probably due to the algorithm used for extracting the cutoff frequency that defines the boundary between low-pass filtered pulses and high-pass filtered noise. We intend to look further at

those sentences in which the margin of superiority of the MIX system was less than expected, to see if their quality can be improved by hand-adjustment of the cutoff frequency. We should stress that the method we used for modifying our model, to produce the five different versions, was extremely crude, since it changed the cutoff frequency through the whole sentence, when a change in part, perhaps a small part, was probably all that was needed.

5.4.4 Buzziness and Breathiness

As in the quality judgments, analysis of variance showed that the main effects of speaker, sentence, and system were all statistically highly significant ($P<0.001$), as were all two-way and three-way interactions between them. The mean judgments of buzziness and breathiness are shown in Figure 2. The absolute levels of the three sets of ratings are not comparable, since no attempt was made to relate the amount of buzziness rated "3" to the amount of breathiness rated "3", or either of these to the quality rated "3". The results of pairwise t-tests between systems for both buzziness and breathiness are shown in Table 3. The BIN system was judged more buzzy than any MIX system ($P<<0.001$), and less breathy than MIX-2, MIX-1 ($P<<0.001$), and MIX-0 ($P<0.001$), and than MIX+1 ($P<0.05$). However, we should stress that the breathiness was only rarely noticeable enough to be objectionable, unlike the buzziness, and these few occasions were all with system MIX-2.

It can be seen from Figure 2 that as the cutoff is moved up in frequency, by moving from MIX-2 to MIX+2, the speech becomes more buzzy and less breathy. The negative correlation between these two trends is statistically significant, although the relationship is not very strong at the level of individual judgments by individual subjects. A confounding factor that may have reduced the magnitude of this relationship is that there were two sources of breathiness in the experiment: one was the manipulation of the MIX source model, and the other was the varying amount of breathiness in the speakers' voices. One of the female speakers, in particular, used a very breathy voice.

Conclusions

The new mixed source model seems to result in appreciably better speech quality than the conventional binary model. It further appears that any remaining weaknesses can probably be ascribed to the algorithm for extracting the cutoff frequency of the filters rather than to the model itself. We need to confirm these results both by hand editing the cutoff frequency in those sentences presently not well handled by the algorithm, and also by using the new source model in conjunction with quantizing the predictor coefficients, to make sure there is no unexpected interaction.

6. REFERENCES

[1] Makhoul, J., and Viswanathan, R., "Adaptive Lattice Methods for Linear Prediction," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978.

[2] Makhoul, J., Viswanathan, R., Schwartz, R., and Huggins, A. W. F., "A mixed source model for Speech Compression and Synthesis." J. Acous. Soc. Am., Vol. 64, pp. 1577-1581, 1978.